

# The FHWA Travel Model Improvement Program Workshop over the Web

The Travel Model Development Series:  
Part I – Travel Model Estimation

---

*presented by*  
Thomas Rossi  
Yasasvi Popuri  
Cambridge Systematics, Inc.

February 10, 2009

1

**Key Message: Purpose of the Webinar Series**

**Details:**

Welcome to the FHWA TMIP Workshop over the Web. This workshop is targeted at Transportation modelers who have a low to moderate level of familiarity with the estimation and validation of travel models.

This series of webinars will introduce the development of model estimation data sets, the structures of the various model components, and the procedures for estimating models. The workshop will include lectures, discussion, and “homework,” that participants will be expected to complete between sessions.

## Webinar Outline

- Session 1: Introduction – October 16, 2008
- Session 2: Data Set Preparation – November 6, 2008
- Session 3: Estimation of Non-Logit Models – December 11, 2008
- Session 4: Estimation of Logit Models – February 10, 2009

2

### **Key Message: Current Session**

#### **Details:**

This session deals with the estimation of logit models, the typical model types and data sources.

Session 5, which will be conducted on March 12, 2009 will cover the various aspects of logit model application.

## Webinar Outline (continued)

- Session 5: Application and Validation of Logit Models – March 12, 2009
- Session 6: Advanced Topics in Discrete Choice Models – April 14, 2009
- Session 7: Trip Assignment – May 7, 2009
- Session 8: Evaluation of Validation Results – June 9, 2009

3

### **Key Message: Upcoming Sessions**

#### **Details:**

Session 5 will be conducted on March 12, 2009; Session 6 will be conducted on April 14, Session 7 on May 7, and Session 8 on June 9.

# Homework

From Session 3

4

**Key Message: Homework Discussion**

**Details:**

Please refer to the homework solutions posted at the website.

## The Use of Logit Models in Transportation Planning

- Can be used to analyze any choice made by travelers with discrete alternatives
- Mode choice is the most common application for which logit models are used in transportation planning
- But there are many other choice processes for which logit models serve well

5

### **Key Message: Logit Model Uses**

#### **Details:**

In this session we will discuss the use of logit models in travel demand modeling. Logit models are used to model any choice with discrete alternatives (modes, number of vehicles available, destination zones, etc.). While mode choice is by far the most common application for logit models in travel modeling, they can be used in many other types of models.

## How Logit Models Work

- Define set of alternatives for the choice
- Define and compute the “utility” of each alternative for each choice maker
- Compute probability of choosing each alternative
- Determine choice share for each alternative

6

### **Key Message: How Logit Models Work**

#### **Details:**

First, a mutually exclusive, collectively exhaustive set of alternatives for the choice is defined.

The “utility” of each alternative is defined and computed for each choice maker. (We will define the term “utility” shortly.)

The probability of choosing each alternative based on its utility relative to other alternatives is computed.

The choice share for each alternative is determined:

By simulating the choice for each individual and summing over all individuals; or

By using the probabilities as shares for specific market segments

## Utility

- Defined as the attractiveness of an alternative based on its attributes and other variables
- Expressed as a function (usually linear)

7

### **Key Message: What is Utility?**

#### **Details:**

Utility is defined as the attractiveness of an alternative expressed by a vector of attributes of the alternative, the decision maker, and the context of the decision.

It defines a single objective function (utility function) expressing the attraction of an alternative in terms of its attributes. For our purposes, this will be a linear function as will be shown shortly.

## Formulation of the Random Utility Model

$$U_i = V_i + e_i$$

where:

$U_i$  = Utility of alternative  $i$

$V_i$  = Deterministic component of  $U_i$

$e_i$  = Random component of  $U_i$

$P(i) = P [U_i > U_j \text{ for all } j \text{ in the choice set}]$

8

### Key Message: Mathematical Formulation for Random Utility

#### Details:

You may recall that in the last session we discussed how we do not know the true values of parameters in travel models. With utilities, not only do we not know the values of the parameters in the utility functions, but the value of the utility itself is unknown. There are things that affect utility that we cannot know or measure. We refer to the part of utility that we can measure and compute as the deterministic component of utility. The portion of utility we cannot measure is referred to as the random component. The “true” utility is the sum of these two components.

A person chooses an alternative if it has a higher utility than any other available alternative. Since we do not know the true value of utility, we make assumptions about the probability distribution of the random component. Using these assumptions, we can compute the probability of each utility being higher than the utilities of all other alternatives, and therefore being the chosen alternative.

## Utility Function

$$U_i = B_{0i} + B_{1i} X_{1i} + B_{2i} X_{2i} + \dots + B_{ni} X_{ni} + e_i$$

where:

$B_{ki}$  = coefficient for variable  $X_{ki}$  for alternative  $i$

$X_{ki}$  = variable that explains choice for alternative  $i$

9

### Key Message: Utility Function

#### Details:

This is the typical linear utility function used in logit models. It is a linear combination of the variables that are known to affect the choice being analyzed and can be quantified, plus the random component, shown here as  $e_i$ .

## Alternative Specific Constants

- Intended to estimate the effects of things that might favor one alternative over another but cannot be quantified
- Assume one constant per alternative, with one assumed to be zero (called the “base alternative”)

10

### Key Message: Alternative Specific Constants

#### Details:

The term  $B_{0i}$  is called the alternative specific constant for alternative  $i$ . It represents the effects of things that might favor one alternative over another but cannot be quantified.

We assume that each alternative has a constant term, with one of the constants assumed to be zero (called the “base alternative”).

## Indicator (“Dummy”) Variables

- Have one of two values, zero or one
- Indicate whether characteristic of traveler, trip, or context is true
- Examples: Income level of household, vehicle availability, peak period

11

### **Key Message: Dummy Variables**

#### **Details:**

Some of the variables ( $X_{ki}$ ) may have a value that is equal to either zero or one, depending on whether a certain condition is met, with no other value possible. They could indicate whether a traveler has particular characteristic, or whether his household does, or his trip does.

## Example Mode Choice Model

- Alternatives are modes (typically 2 to ~25)
- Utilities represent characteristics of modes, travelers, and context
- Choice probabilities used to determine mode shares

12

### **Key Message: Mode Choice Model Example**

#### **Details:**

Consider the example of a mode choice model:

The alternatives in a mode choice are the modes. Typically there could be anywhere from 2 to over 25 modes. They might include drive alone, shared ride, transit-auto access, transit-walk access, walk, bicycle, etc.

The utilities represent characteristics of the modes, travelers, and context.

Characteristics of modes: travel time, cost, transfers

Characteristics of travelers: income, auto ownership

Context: area type

The choice probabilities, computed from the utilities, are used to determine the mode shares.

## Example Destination Choice (Trip Distribution)

- Alternatives are zones
- Utilities represent travel impedance and characteristics of destinations, travelers, and context
- Choice probabilities used to determine shares of trips to each zone

13

### **Key Message: Destination Choice Model Example**

#### **Details:**

Consider another example – the destination choice model, often used for trip distribution:

The alternatives are usually the zones for the attraction ends of the trips.

The utilities represent the impedance of travel to the destination and the characteristics of destinations, travelers, and context. For example, variables could include travel time, cost, transfers, income, and area type.

The choice probabilities are used to determine the shares of trips from each zone to each other zone.

## Example Vehicle Availability Model

- Alternatives are number of vehicles owned (0, 1, 2, 3, etc.)
- Utilities represent characteristics of households, areas, and context
- Choice probabilities used to determine shares of households for each vehicle availability level by zone

14

### **Key Message: Vehicle Availability Model Example**

#### **Details:**

Another example of a logit model used in many travel models is the vehicle availability (auto ownership) model:

The alternatives are the number of vehicles owned by a household (0, 1, 2, 3, etc.)

The utilities represent the characteristics of households, areas (zones), and context. For example, variables could include the number of persons, number of workers, and income of the household, and the zone's area type.

The choice probabilities are used to determine shares of households by zone with 0 vehicles, 1 vehicle, 2 vehicles, etc.

## Other Examples

- Time of day choice models
- Socioeconomic models
- Trip generation
- Land use models
- Components of activity based models

15

### **Key Message: Other Examples**

#### **Details:**

There are many other examples where logit models have been used as components of travel demand models:

Time of day choice models have been estimated where the time of day in which a trip is made is estimated depending on trip characteristics, congestion levels, etc.

Socioeconomic models, such as number of workers per household, number of children, etc. have been estimated, similar to vehicle availability models.

While not common, even trip generation can be modeled using a logit model, where the choices include making zero trips, one trip, etc.

Some integrated land use-transportation models use logit formulations.

In most activity based models, various components use logit formulations. All current models used in U.S. by planning agencies are series of logit models

Daily activity patterns

Destination, mode, and time of day choice at tour and trip level

Other choices (e.g., auto ownership)

Models are applied disaggregately by simulating the choices of each individual.

## The Binary Logit Formulation

$$P_r(1) = \frac{\exp(v_1)}{\exp(v_1) + \exp(v_2)}$$

$$P_r(1) = \frac{1}{1 + \exp[-(v_1 - v_2)]}$$

Only the differences in utilities matter

16

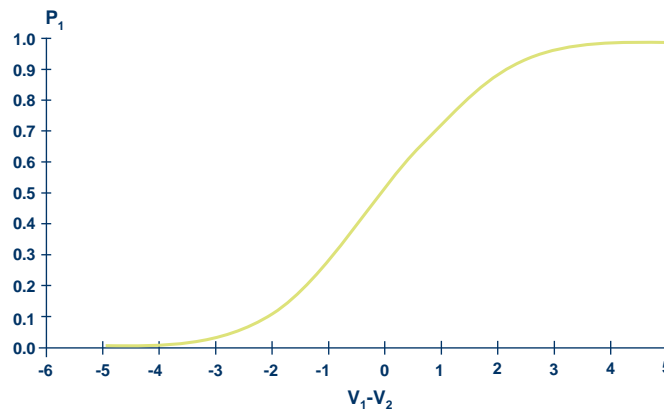
### Key Message: Logit Formulation

#### Details:

The simplest type of logit model is the binary model, where there are only two alternatives to choose from. The probability of choosing an alternative is a function of the deterministic components of the utilities of the two alternatives. (For simplicity, I will call the deterministic components “utilities” from here on.) The utilities are exponentiated (the number  $e$  take to the power of the utility), and the probability of each alternative is the exponentiated utility divided by the sum of the two exponentiated utilities. It is easy to see that the sum of the two probabilities must be one, that the probabilities must be between zero and one, and that if the utilities are finite numbers, neither alternative has a probability of exactly zero or one (although they can be very close).

The probability of an alternative can also be expressed as a function of the differences between the two utilities. So only the difference between the two utilities affects the choice; it doesn't matter what the actual values are. The probability of choosing an alternative is the same whether the two utilities are 0 and 2, or 100 and 102.

## The Binary Logit Model



17

### Key Message: The Binary Logit Model

#### Details:

This is a graph of the probability of choosing alternative 1 as a function of the difference between the two utilities. If  $V_1$  is greater than  $V_2$ , the probability of choosing alternative 1 is greater than 0.5. If the two utilities are equal, the probability of each alternative is 0.5. If  $V_1$  is much greater than  $V_2$ , the probability of alternative 1 gets close to 1.0. If  $V_2$  is much greater than  $V_1$ , the probability of alternative 1 gets close to zero.

(You may have heard some people refer to this as the “S-shaped” logit curve.)

## The Multinomial Logit Model

**Extending the binary model to three modes:**

$$P(1) = \frac{\exp(v_1)}{\exp(v_1) + \exp(v_2) + \exp(v_3)}$$

**Extending to n modes:**

$$P(1) = \frac{\exp(v_1)}{\exp(v_1) + \exp(v_2) + \dots + \exp(v_n)}$$

18

### **Key Message: The Multinomial Logit Model**

#### **Details:**

The multinomial logit model extends the binary model to more than two alternatives. The probability of choosing an alternative is a function of the deterministic components of the utilities of all alternatives. Again, the utilities are exponentiated, and the probability of each alternative is the exponentiated utility divided by the sum of the exponentiated utilities for all alternatives. Again, the sum of the probabilities must be one and the probabilities must be between zero and one, and if the utilities are finite numbers, no alternative has a probability of exactly zero or one (although they can be very close).

## MNL Model Properties

- $P_i$ , all  $i$  depends on the deterministic components of the utilities of all alternatives ( $V_j$ , all  $j$ )
- $P_i$  increases as  $V_i$  increases, and decreases as  $V_j$  ( $j \neq i$ ) increases
- $P_i$  depends only on the values  $(V_j - V_i)$  for all alternatives  $j$ , except  $i$

19

### Key Message: The Multinomial Logit Model Properties

#### Details:

Each probability depends on the deterministic components of the utilities of all alternatives.

Each probability increases as its utility increases, and decreases as the utility of any other alternative increases.

The probability that alternative  $i$  is chosen depends only on the differences between its utility and that of each other alternative, that is  $(V_j - V_i)$  for all alternatives  $j$ , except  $i$  itself.

The MNL can be used for any number of alternatives and is relatively easy to understand and apply.

## The IIA Property Definition

- The independence from irrelevant alternatives property  
“For any individual, the ratio of the probabilities of choosing two (available) alternatives is independent of the availability or attributes of any other alternative”
- Mathematically

$$\frac{\Pr(i)}{\Pr(k)} = \frac{\exp(V_i)}{\exp(V_k)} = \exp(V_i - V_k)$$

20

### Key Message: The Independence of Irrelevant Alternatives Property

#### Details:

An important property of multinomial logit models is independence from irrelevant alternatives, or IIA. “For any individual, the ratio of the probabilities of choosing two (available) alternatives is independent of the availability or attributes of any other alternative.”

The effects of the IIA property can be shown through an example. Let's say that the current probabilities of choosing auto, bus, and rail are 60%, 20%, and 20% respectively. Now let's say that rail improvements are made so that the new rail share is 40%. This means that, since the ratio of the auto and bus probabilities is 3, the new probabilities must be 45% auto, 15% bus, and 40% rail. The probability of switching to rail is the same for a bus rider or an auto traveler. If the rail is essentially competing with the bus, this would not make sense.

## The IIA Property Red and Blue

- Scenario 1

- Available modes are auto (da) and red buses (rb); red buses have plenty of seats for all passengers
- $V_{da} = V_{rb}$
- MNL model says  $\Pr(da) = \Pr(rb) = 0.5$

- Scenario 2

- A new bus operator exactly duplicates red bus service using blue buses (bb)
- MNL model says  $\Pr(da) = \Pr(rb) = \Pr(bb) = 0.33$

21

### Key Message: The Independence of Irrelevant Alternatives Property

#### Details:

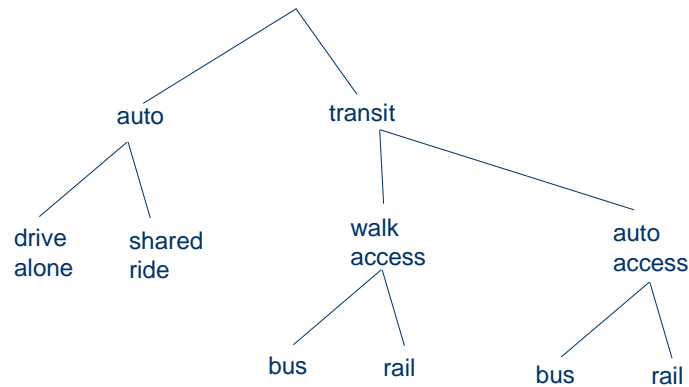
One way of looking at the IIA property's effects is through the classic “red bus, blue bus” paradox.

Say that at present, the available modes are auto and red buses, and red buses have plenty of seats for all passengers. Let's say that currently the utilities of auto and red bus are equal, and so each has a 50% probability.

Now, say a new bus operator decides to exactly duplicate the red bus service using blue buses. If “blue bus” is considered a new mode, then it will have the same utility as red bus (since the service is the same), and so the probabilities of the three modes would be 1/3. In reality, if the two bus services were perceived the same, the probabilities would be 50% auto, 25% red bus, and 25% blue bus.

While we would not be so naïve to define mode alternatives based on vehicle color, there are many cases where one alternative is similar to another, more than it is similar to the remaining modes.

## The Nested Logit Model



22

### Key Message: The Nested Logit Model

#### Details:

The IIA property of MNL is one motivation for the use of nested logit models. In a nested model, alternatives that are more similar to one another are grouped into “nests.” The probability of each alternative is the probability of the alternative within its nest, conditional on the probability of the nest (the upper level alternative) being chosen. For example, in the diagram shown here, the probability of drive alone is the probability of drive alone given that auto is chosen, conditional on the probability that auto is chosen (instead of transit, the only other alternative at the upper level).

## Utilities in Nested Logit

- Logsum variable
  - $\text{Logsum} = \ln \sum [\exp (V_i)]$  for all alts.  $i$  in the nest
- Utility for nest
  - $V_{\text{nest}} = B_{0(\text{nest})} + B_{\text{logsum}} \text{logsum}$
  - $0 < B_{\text{logsum}} < 1$

23

### Key Message: The Nested Logit Model Utilities

#### Details:

The utilities (and therefore probabilities) of alternatives in a nested model are computed using a variable called a logsum. The logsum is defined as the natural logarithm of the sum of the exponentiated utilities for all alternatives in the nest. If the nest was considered its own multinomial logit model, it would be the log of the denominator of the corresponding probability equations.

The utility of the nest is computed as the constant for the combined alternative represented by the nest plus the logsum variable multiplied by its coefficient. This constant and the logsum coefficient can be estimated the same way other utility parameters are estimated, using statistical software.

A valid logsum coefficient must be between zero and one.

## Choosing the Independent Variables in Utility Functions

- Relevance to the travel choice
- Availability in the estimation data set
- Availability for model application (forecasting)
- Statistical testing

24

### **Key Message: Independent Variables in the Utility Functions**

#### **Details:**

We saw a slide very similar to this in an earlier webinar session. This is a reminder how we choose the variables to include in a model. In the case of the logit model, we refer to variables that will be in the utility function.

The independent variables are those on which the travel choice is based. The important factors in choosing the independent variables include:

- The relevance of the variables in explaining the travel choice being modeled.
- The availability of the variables in the data set. For example, if income was not asked in a survey, it is not available to be used in models estimated from that survey data set.
- The availability of the variable for forecasting. For example, if income is asked in the survey but income forecasts are not available, it cannot be used.
- The results of statistical tests that demonstrate that the variable helps explain the travel choice.

## Setting up Model Estimation Data Set

- Survey data organized by individual records
  - For example, for mode choice, each record is a trip
- Additional variables added to data set
  - Information on transportation level of service (from skims)
  - Information at zone level, such as parking costs (from lookup tables)
  - Transformations of variables, as needed

25

### **Key Message: Data Preparation for Model Estimation**

#### **Details:**

The setup of the model estimation data set is similar to what we have seen for other types of models, such as regression models. We start with the relevant file from the survey data set. For mode choice, for example, we would use trip records from household and transit on-board survey data sets.

As we discussed in an earlier session, we use transportation level of service data from the model's transportation networks, based on the observed trip origin and destination. This is often referred to as "attaching the skim data" to the survey data set. We may also attach other data, such as parking cost data by zone, to the data set.

## Setting up Model Estimation Process

- Define utility functions
- Prepare input files for model estimation software

26

### **Key Message: Data Preparation for Model Estimation**

#### **Details:**

Next we define the actual utility functions, and then we prepare the input files for model estimation. The software determines the format of these files, but they must include the definition of the choice, the alternatives, the utility functions, and the parameters to be estimated.

## Trip Data File Typical Fields

### From the survey

- Origin and destination
- Trip purpose
- Chosen mode
- Time of day of trip
- **Trip time/cost**
- Household/person characteristics (linked from household/person file)

### From other sources

- Travel time (in-vehicle)
- Other time components (wait, access/egress, transfer)
- Costs (parking, auto operating, transit fare)
- Number of transit transfers
- Zone attributes
- Logsums from other models

27

### Key Message: Trip Data File Typical Fields

#### Details:

Review of typical fields in the trip data file. We have seen this before.

## Example Data Set

### Mode Choice Model Estimation (Partial)

Trip ID	Trip Purpose	Chosen Mode	Origin	Dest	IVT	Transit Walk Access		
						Wait	Walk	Fare
10001	1	1	107	12	20.1	10	15	\$1.00
10002	3	1	12	118	10.3	5	2	\$1.00
10003	2	1	118	107	17.2	10	15	\$1.00

28

#### Key Message: Model Estimation Data Set Example

##### Details:

Here is an example of part of a mode choice model estimation set developed from the trip files from household/transit on-board surveys. Trip purpose is an important field since separate mode choice parameters are estimated for different trip purposes. The origin and destination zones are used to determine the values of variables that come from other data sources. For example, the transportation level of service data comes from the network skims, based on the origin and destination (or production and attraction) zones. Shown here are some level of service variables for a single mode, transit with walk access. The data set must include values for these variables for every mode alternative in the model. Also included are other characteristics of the traveler and his/her household that might be relevant to the choice, and other variables that might be rated to the origin or destination (e.g. parking cost).

## Estimating the Logit Model

1. Run statistical software to estimate coefficients of utility functions
2. Evaluate results
3. Revise specification and reestimate
  - Consider alternative variable definitions, combinations
  - Eliminate variables as appropriate
  - Consider constraining parameter estimates, but only when necessary
4. Choose “best” specification

29

### **Key Message: Logit Model Estimation Details**

#### **Details:**

It is relatively quick to run statistical software to obtain parameter estimates for most simple logit models. The results are evaluated, and changes made to the model to try to improve the results. These changes could include revising, combining, or eliminating variables.

It is also important to note that unreasonable results can be due to data problems. These results may indicate the need to recheck the data set.

When everything the modelers think of is exhausted, the best specification can be determined.

## Constraining Coefficient Estimates

- Inequality constraints, e.g.  $B_k \leq 0$
- Fixed value constraints, e.g.  $B_k = C$
- Linear constraints, e.g.  $B_k = C B_j$ 
  - Assumed value of time:  $B_{\text{vt}} = (\text{VOT}) B_{\text{cost}}$

30

### Key Message: Constraining the Parameters of a Logit Model

#### Details:

We would greatly prefer that the estimated values of parameters be reasonable and ready to use, but this is not always the case. It is sometimes desirable to constrain parameter values to certain values or ranges. These may take the form of inequality constraints, such as constraining a value to be less than or equal to zero, constraining a parameter estimate to equal a specific value, or constraining the value of one parameter relative to another. A common example of the latter is constraining the coefficient for in-vehicle time for one mode to be equal to that of another mode. Another example is constraining the implied “value of time,” which is the relationship between the time and cost coefficients in a mode choice model.

## Constraining Coefficient Estimates Why Do It?

- Estimated values not reasonable in sign or magnitude (e.g.  $B_{ivt} > 0$ )
- Estimate of parameter value unreasonable relative to others (e.g.  $|B_{ivt}| > |B_{ovt}|$ )
- Consistency with other sources desired (e.g. FTA New Starts Guidelines)

31

### Key Message: Why Constrain the Parameters of a Logit Model

#### Details:

Why would we want to constrain parameter values to something different than the model estimation process tells us?

The main issue is that the estimated values are illogical or unreasonable compared to experience with other models. For example, the coefficients of time and cost variables in a mode choice model should be negative. Otherwise, the model would result in increased demand for a mode that is made worse (longer in time or higher in cost).

Sometimes when parameter estimates are individually reasonable, they are not reasonable when viewed together. For example, in a mode choice model, the coefficients for out-of-vehicle time variables should be greater in absolute value than those for in-vehicle time variables. Travelers generally would prefer to spend time traveling in an vehicle than the same amount of time waiting or walking to the vehicle.

There may also be reasons for model parameters to be consistent with other data sources. For example, FTA has guidelines for parameters of mode choice models used in ridership forecasting for New Starts projects.

## Evaluating the Logit Model Estimation Results

1. Reasonableness of coefficient estimates
  - Sign
  - Magnitudes
2. Significance of estimates (t-statistics)
  - $|t| > 1.96$  implies significance at 95% level
3. Goodness of fit ( $\rho^2$ )

32

### Key Message: Evaluating the Logit Model Results

#### Details:

How do we check the model estimation results for reasonableness?

The signs and magnitudes of coefficients must be checked for reasonableness. If a variable is associated with an alternative being more attractive, then it should have a positive sign since an increase in the value of the variable would increase the utility. As we just discussed in the case of level of service variables in mode choice models, some variables are inversely related to an alternative's utility, and so the signs of coefficients should be negative.

Magnitudes and relative magnitudes of variables should also be checked. The magnitude of a coefficient is related to the sensitivity of the choice to the variable. In some cases, results from other models can be compared, but in other cases, the sensitivity cannot be estimated until validation is done.

Statistical significance of the coefficient estimates is tested using the t-statistics, as we saw in the session where linear regression models were discussed. If coefficient estimates make sense, they are often retained even if the significance level is lower than 90%.

The  $\rho^2$  value measures the goodness of fit for the entire model and is computed from the likelihood ratios.  $\rho^2$  can be measured with respect to zero or to constants; often, both measures are reported and analyzed. The  $\rho^2$  value by itself can be hard to interpret although it should be well above zero to show that variables other than the constants have a significant effect on utility.  $\rho^2$  can be quite valuable in comparing alternative model formulations; a higher  $\rho^2$ , adjusted for the degrees of freedom (based on the number of variables), means a superior model formulation.

## Vehicle Availability Model Example Estimation Results

Variable	Vehicle Availability Level				
	0	1	2	3	4+
Persons per household	--	--	0.1164 (2.1)	0.1164 (2.1)	0.2571 (2.1)
Workers per household	--	--	0.4915 (5.2)	1.474 (10.8)	2.139 (10.0)
Household density	--	-0.0458 (-2.9)	-0.1327 (-5.4)	-0.1717 (-4.4)	-0.2549 (-3.0)
ln(income)	--	1.130 (8.7)	2.497 (13.9)	2.995 (12.7)	3.242 (7.6)
Transit/highway accessibility	--	-1.133 (-1.7)	-2.054 (-2.8)	-2.742 (-3.3)	-2.742 (-3.3)
Persons less than vehicles	--	--	-2.870 (-8.8)	-1.017 (-5.3)	-0.5181 (1.1)
Constant	--	0.164 (0.2)	-3.761 (-4.6)	-8.229 (-8.0)	-12.87 (6.8)

33

### Key Message: Vehicle Availability Model Estimation Results

#### Details:

This is an example of estimation results for a multinomial logit model of vehicle availability. The alternatives are 0, 1, 2, 3, and 4+ vehicles available to households.

Logit model results are often presented in tables like this. The utility of each alternative is read downward. For example, the utility of one vehicle is:

$-0.0458 * \text{household density} + 1.130 * \ln(\text{income}) - 0.4277 * \text{pedestrian environment variable} - 1.133 * \text{transit/highway access ratio} + 0.164$ .

Just to define some variables:

- Household density is the number of households per acre in the home zone
- ln (household income) is the natural log of the household income in thousands of dollars
- Transit/highway access ratio is the ratio of the percentage of regional employment that can be reached within 80 minutes using transit divided by the percentage of employment that can be reached within 60 minutes using highway
- Persons less than vehicles = 1 if the number of persons in the household is less than the number of vehicles, zero otherwise

## Vehicle Availability Model Example Estimation Results

Variable	Vehicle Availability Level				
	0	1	2	3	4+
Persons per household	--	--	0.1164 (2.1)	0.1164 (2.1)	0.2571 (2.1)
Workers per household	--	--	0.4915 (5.2)	1.474 (10.8)	2.139 (10.0)
Household density	--	-0.0458 (-2.9)	-0.1327 (-5.4)	-0.1717 (-4.4)	-0.2549 (-3.0)
ln(income)	--	1.130 (8.7)	2.497 (13.9)	2.995 (12.7)	3.242 (7.6)
Transit/highway accessibility	--	-1.133 (-1.7)	-2.054 (-2.8)	-2.742 (-3.3)	-2.742 (-3.3)
Persons less than vehicles	--	--	-2.870 (-8.8)	-1.017 (-5.3)	-0.5181 (1.1)
Constant	--	0.164 (0.2)	-3.761 (-4.6)	-8.229 (-8.0)	-12.87 (6.8)
$\rho^2$ w.r.t zero = 0.447		$\rho^2$ w.r.t constants = 0.302			

34

### Key Message: Vehicle Availability Model Estimation Results

#### Details:

Let's interpret the results of this model:

- Note that zero vehicles is the "base alternative" and has a utility of zero (all coefficients constrained to zero).
- As the number of persons or workers in a household or the household income increases, the more likely that the number of vehicles increase. The coefficients for these variables become more positive for higher numbers of vehicles.
- An additional worker has more effect on the number of vehicles owned than an additional non-worker in the household. The coefficient on workers is greater in all cases.
- As residential density or transit accessibility relative to highway accessibility increases, the likelihood of having more vehicles decreases. The coefficients for these variables become more positive for higher numbers of vehicles.
- There is a decreased utility of having more vehicles than persons in the household. For example, the utility of having two vehicles is significantly reduced if there is only one person in the household.
- The t-statistics show that all estimates except two are significant at the 90% level. These could be removed from the model although the persons < vehicles for 4+ vehicles is of reasonable sign and magnitude.
- The  $\rho^2$  values with respect to zero and constants are well above zero.

## Mode Choice Model Example Estimation Results

Variable	Drive Alone	Shared Ride	Walk	Transit Auto Access	Transit Walk Access
In-vehicle time (min)	-0.020 (-2.7)	-0.020 (-2.7)	--	-0.020 (-2.7)	-0.020 (-2.7)
Out-of-vehicle time (min)	--	--	--	-0.045 (-3.7)	-0.045 (-3.7)
Cost (\$)	-0.400 (-7.0)	-0.400 (-7.0)	--	-0.400 (-7.0)	-0.400 (-7.0)
Distance (miles)	--	--	-1.63 (-7.2)	--	--
Zero vehicles owned	--	--	1.18 (1.7)	1.16 (1.8)	2.49 (4.6)
Vehicles/person	--	-2.45 (-10.0)	-6.87 (-6.2)	-5.20 (-6.8)	-5.20 (-6.8)
Population density (1000/acre)	--	--	0.042 (3.7)	0.030 (3.1)	0.030 (3.1)
Employment density (1000/acre)	--	--	0.31 (4.2)	0.079 (1.2)	0.079 (1.2)
Constant	--	1.65 (8.3)	3.75 (5.9)	3.55 (6.5)	6.83 (14.2)
$\rho^2$ w.r.t zero = 0.319		$\rho^2$ w.r.t constants = 0.156			

35

### Key Message: Mode Choice Model Estimation Results

#### Details:

These are results from a multinomial logit mode choice model estimation for home based shopping trips.

- The in-vehicle time coefficient is consistent with the experience from other models. The ratio of the out-of-vehicle and in-vehicle time coefficients is 2.5, also consistent with other models.
- The implied value of time is computed as the ratio of the in-vehicle time and cost coefficients. This equals  $-0.020/-0.400 = \$0.05/\text{minute} = \$3.00/\text{hour}$ . This seems reasonable and is consistent with other model results.
- Households with zero vehicles are more likely to use transit or walk (positive coefficients for these modes). Owning more cars per person in the household makes it less likely that shared ride will be chosen, even less likely for transit, and even less so for walking.
- Higher development density increases the likelihood of transit use and increases it further for walking.
- The t-statistics show that all estimates except for employment density for transit are significant at the 90% level. The coefficient is of reasonable sign and magnitude, though.
- The  $\rho^2$  values with respect to zero and constants are well above zero.

## Typical Selected Home-Based Work Parameters

	Average from U.S. Cities	Typical Range
In-vehicle time	-0.028	-0.01 to -0.05
Out-of-vehicle time	-0.054	-0.03 to -0.07
Cost	-0.720	-0.2 to -1.3
Ratio: $B_{ovt} / B_{ivt}$	1.9	1.5 to 3.0
Value of Time	\$2.30	\$2.00 to \$5.00

36

### Key Message: Typical Parameters for Home-Based Work Trips

#### Details:

I had mentioned that some of the mode choice parameters were consistent with other models. This slide shows the averages and ranges of some parameters that have been estimated in mode choice models throughout the U.S. These can be used for some “reality checks” of the estimated coefficients.

But these numbers should be used with caution, because nearly all mode choice models include several variables other than those listed, and the presence or absence of other variables can affect the parameter values. Unless mode specifications are identical, the parameters should not be expected to be identical.

Furthermore, there are differences between urban areas. The sensitivity of mode choice to, say, walking time may depend on factors such as climate.

In addition, costs (and therefore values of time) can depend on what year the data represent. The average value of time of \$2.30 represents data collected in the 1980s and 1990s.

## Estimation Problems

- Use of too many alternative-specific constants
- Constraining coefficients across alternatives for non-varying variables
- Perfect collinearity of variables

37

### Key Message: Estimation Problems

#### Details:

This slide lists some common logit model estimation problems that may occur (the model estimation will likely crash!).

First, it should always be remembered that each alternative in a multinomial logit model can have a constant except one. It does not matter which alternative is chosen to have a value zero, but one of them must—always! Since only the differences between utilities matter in computing choice probabilities, the model estimation software must have a “base” from which to estimate everything else relatively. In a nested model, this is true for each level of the nest.

Each variable must be able to take on different values for all alternatives for which the same coefficient will be estimated. For example, if the variable “number of autos” is used in a mode choice model for all modes, a coefficient cannot be estimated, since the number of autos is not different for each alternative. It could be used with different coefficients for each alternative, except one (again, a base is needed). A variable such as “in-vehicle time” can be used with the same coefficient for all alternatives because the in-vehicle time is not the same for each mode (IVT can be considered separate variables for each mode).

No two variables can be perfectly collinear—that is, one variable cannot be exactly proportional to another. For example, one couldn’t use both time and distance for the walk mode in a mode choice model if the same walking speed is assumed for everyone. The model estimation process could not separate the effects of the two variables.

# Homework

## Session 4

38

### **Key Message: Homework**

#### **Details:**

The homework for Session 4 can be downloaded from the course website. We would strongly recommend that the participants work through the homework problems to get more value out of this session.